that one expects to find in a single 16S rRNA sequence. Thus, the hepta-mers (4^7 = 16,384 in total) represent the smallest sequence length that is likely to produce meaningful signature information. On the opposite side, large oligonucleotides tend to be unique to individual organisms. That is to say, as oligonucleotide size increases, a larger portion of the signatures will be for leaf nodes, e.g. small numbers of closely related organisms and a decreasing percentage will signify internal nodes. Based on prior experience with 16S rRNA ribonuclease T1 oligonucleotides, it is likely that sequences larger than length 15 will mainly have utility for leaf nodes.

Design and implementations

Programming language

Except the first program readseq, which is preinstalled as a binary executable, all other programs developed for this project were written in Perl.

Perl is a freely available, non-proprietary, open-source programming language. Thus, programs written in Perl will not be affected by possible future changes in the license of the language compiler/interpreter. Perl is also a very high-level language for general purposes. It has 4 function points per 100 lines of code, compared with 0.8 for C and 2 for C++. This means that software development in Perl is generally much faster than that in most other programming languages. Perl is especially efficient in dealing with text, which makes it an appropriate choice for manipulating genetic sequences. In addition, Perl's excellent built-in data structures, automatic garbage collection, and almost unrivalled portability also make it more attractive.

More information on Perl and its current releases can be found on the Perl web site: http://www.perl.com/2.1

Data structures

All Perl built-in data structures, namely scalar, array, and hash, are used in this invention. Because of the complexity of the data presentations, more sophisticated data structures such as bi-directional binary tree and composite hash, are also used.

Given the characteristic structure of the phylogenetic tree, it was natural to represent it as a binary tree in the program. In this case the tree structure is special in that it is bi-directional. The parent tree node has a pointer to each of its two child tree nodes and the child tree node also has a pointer back to its parent tree node (Figure 3). This unusual tree structure is required to facilitate the signature quality index value calculation at each branch tree node (excluding the tree root and all the leaf nodes).

Each leaf tree node has five data fields: "shortName", "fullName", "leafNumber", "isValid", and "isMatched" (Figure 1). The first two fields hold the abbreviated name and the full name of the prokaryote.

REPLACED PAGE

that one expects to find in a single 16S rRNA sequence. Thus, the heptamers ($4^7 = 16{,}384$ in total) represent the smallest sequence length that is likely to produce meaningful signature information. On the opposite side, large oligonucleotides tend to be unique to individual organisms. That is to say, as oligonucleotide size increases, a larger portion of the signatures will be for leaf nodes, e.g. small numbers of closely related organisms and a decreasing percentage will signify internal nodes. Based on prior experience with 16S rRNA ribonuclease T1 oligonucleotides, it is likely that sequences larger than length 15 will mainly have utility for leaf nodes.

Design and implementations

Programming language

Except the first program readseq, which is preinstalled as a binary executable, all other programs developed for this project were written in Perl.

Perl is a freely available, non-proprietary, open-source programming language. Thus, programs written in Perl will not be affected by possible future changes in the license of the language compiler/interpreter. Perl is also a very high-level language for general purposes. It has 4 function points per 100 lines of code, compared with 0.8 for C and 2 for C++. This means that software development in Perl is generally much faster than that in most other programming languages. Perl is especially efficient in dealing with text, which makes it an appropriate choice for manipulating genetic sequences. In addition, Perl's excellent built-in data structures, automatic garbage collection, and almost unrivalled portability also make it more attractive.

More information on Perl and its newest release can be found at the Perl web site: http://www.perl.com.2.2 Data structures.

All Perl built-in data structures, namely scalar, array, and hash, are used in this invention. Because of the complexity of the data presentations, more sophisticated data structures such as bi-directional binary tree and composite hash, are also used.

Given the characteristic structure of the phylogenetic tree, it was natural to represent it as a binary tree in the program. In this case the tree structure is special in that it is bi-directional. The parent tree node has a pointer to each of its two child tree nodes and the child tree node also has a pointer back to its parent tree node (**Figure 1**). This unusual tree structure is required to facilitate the signature quality index value calculation at each branch tree node (excluding the tree root and all the leaf nodes).

Each leaf tree node has five data fields: "shortName", "fullName", "leafNumber", "isValid", and "isMatched" (**Figure 1**). The first two fields hold the abbreviated name and the full name of the prokaryote.

Respectfully submitted,

Richard Coale Willson, Jr.
Attorney for Applicants
Registration No. 22,080
USPTO Customer 26830
Technology Licensing Co. LIC
3205 Harvest Moon Ste 200
Telephone - 727 781 0089
Fax: 727 785 8435
rwillso@gmail.com

010AUS Signatures Correction of Defect page 12